(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開平10－198683
(43)公開日　平成10年(1998)７月31日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ
Ｇ０６Ｋ　9/62	６１０	Ｇ０６Ｋ　9/62	６１０Ｃ
		Ｇ０６Ｆ　15/40	３７０Ｂ

審査請求　未請求　請求項の数５　ＯＬ　（全　５　頁）

(21)出願番号　特願平9－738

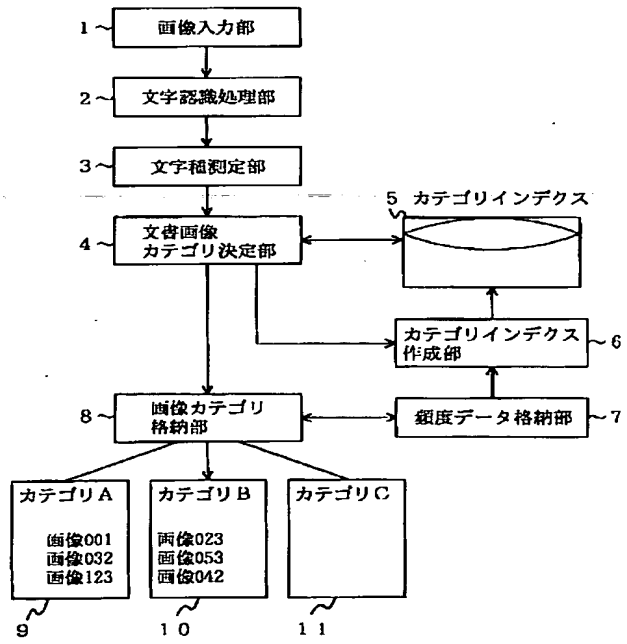
(22)出願日　平成９年(1997)１月７日

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　大阿久　志緒理
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(72)発明者　齋藤　高志
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(72)発明者　阿部　悌
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(74)代理人　弁理士　鈴木　誠　（外１名）

最終頁に続く

(54)【発明の名称】　文書画像分類方法

(57)【要約】
【課題】　小規模なシステム構成で高速に画像分類・検索を行なうために、文字認識結果の文字種情報を基に、帳票・電話帳・日本語一般文書・その他の言語の文書などを識別し、文書画像を自動分類する。
【解決手段】　カテゴリインデクス作成部６は、カテゴリを代表する文書画像の文字種別頻度データを作成してカテゴリインデクス５に格納する。入力文書画像が文字認識され、文字種測定部３では、認識結果から文字種毎の頻度を測定する。カテゴリ決定部４は、測定された頻度データとカテゴリインデクス５の頻度データとの類似度を求め、最も類似度の高い代表画像が属するカテゴリを、入力文書画像のカテゴリと決定する。



【特許請求の範囲】

【請求項１】　複数の文書画像を所定のカテゴリに分類する文書画像分類方法であって、入力された文書画像に対して文字認識処理を行い、認識処理された文字種の特徴を基に前記入力文書画像を所定のカテゴリに分類することを特徴とする文書画像分類方法。

【請求項２】　複数の文書画像を所定のカテゴリに分類する文書画像分類方法であって、入力された文書画像に対して文字認識処理を行い、認識処理された文字種の特徴および総文字数を基に所定の文書画像との類似度を測定し、前記入力文書画像を、最も類似度の高い文書画像と同一のカテゴリに分類することを特徴とする文書画像分類方法。

【請求項３】　複数の入力文書画像を所定のカテゴリに分類する文書画像分類方法であって、予め用意された画像に対して文字認識処理を行い、認識処理された文字種の特徴および総文字数を測定し、前記画像を所定のカテゴリに分類し、該分類された各カテゴリの特徴を最も示す代表画像を選択し、入力された文書画像のカテゴリを決定する際に、文字認識処理を行い、認識処理された文字種の特徴および総文字数を測定し、前記代表画像との類似度を測定し、最も類似度の高い代表画像が所属するカテゴリに決定することを特徴とする文書画像分類方法。

【請求項４】　前記何れの代表画像とも類似しないとき、すべての原稿の文字種データを用いて前記カテゴリを再設定することを特徴とする請求項３記載の文書画像分類方法。

【請求項５】　前記文字種の特徴として、数字または英字の頻度を用いることを特徴とする請求項１、２、３または４記載の文書画像分類方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文字認識結果の文字種情報を基に文書を識別し、文書画像を自動的に分類する文書画像分類方法に関する。

【０００２】

【従来の技術】従来、文書画像の分類・検索においては、ユーザーがキーワードを付与し、あらかじめインデクスを作成し、キーワードとインデクスの内容の照合により分類・検索する方法が採られてきた。

【０００３】また、他の方法としては、インデクスとの照合ではなく、文書画像を文字認識によりテキスト化して格納しておき、その内容とキーワードを照合し、画像を検索するなどの方法も提案されている（例えば、特開平８－７０３３号公報を参照）。

【０００４】

【発明が解決しようとする課題】しかし、上記した方法は何れもユーザーがキーワードをその都度指定しなければならず、繁雑な作業が要求される。また、キーワードの内容や指定の仕方によっては、所望の画像が得られないこともあり、ユーザーに多大な負担を与えてしまう。

【０００５】そこで、文書画像を分類・検索する際に、文字認識を行って文書中のキーワードを自動抽出し、分類する手法が提案されている。例えば、特開平７－１１４５７２号公報に記載された技術では、文書から自動的に単語の特徴ベクトルを抽出し、その特徴ベクトルを基に文書を分類することにより、意味的な異なりを用いた自動分類を実現している。

【０００６】しかし、このような方法は、キーワードの相関関係を記述する手段が複雑になるとともに、シソーラスのような大規模な言語データベースなども必要となり、システム構成が大規模なものとなってしまう。

【０００７】本発明は上記した事情を考慮してなされたもので、本発明の目的は、小規模なシステム構成で高速に画像分類・検索を行なうために、文字認識結果の文字種情報を基に、帳票・電話帳・日本語一般文書・その他の言語の文書などを識別し、文書画像を自動分類する文書画像分類方法を提供することにある。

【０００８】

【課題を解決するための手段】前記目的を達成するために、請求項１記載の発明では、複数の文書画像を所定のカテゴリに分類する文書画像分類方法であって、入力された文書画像に対して文字認識処理を行い、認識処理された文字種の特徴を基に前記入力文書画像を所定のカテゴリに分類することを特徴としている。

【０００９】請求項２記載の発明では、複数の文書画像を所定のカテゴリに分類する文書画像分類方法であって、入力された文書画像に対して文字認識処理を行い、認識処理された文字種の特徴および総文字数を基に所定の文書画像との類似度を測定し、前記入力文書画像を、最も類似度の高い文書画像と同一のカテゴリに分類することを特徴としている。

【００１０】請求項３記載の発明では、複数の入力文書画像を所定のカテゴリに分類する文書画像分類方法であって、予め用意された画像に対して文字認識処理を行い、認識処理された文字種の特徴および総文字数を測定し、前記画像を所定のカテゴリに分類し、該分類された各カテゴリの特徴を最も示す代表画像を選択し、入力された文書画像のカテゴリを決定する際に、文字認識処理を行い、認識処理された文字種の特徴および総文字数を測定し、前記代表画像との類似度を測定し、最も類似度の高い代表画像が所属するカテゴリに決定することを特徴としている。

【００１１】請求項４記載の発明では、前記何れの代表画像とも類似しないとき、すべての原稿の文字種データを用いて前記カテゴリを再設定することを特徴としている。

【００１２】請求項５記載の発明では、前記文字種の特徴として、数字または英字の頻度を用いることを特徴と

している。

【００１３】

【発明の実施の形態】以下、本発明の一実施例を図面を用いて具体的に説明する。図１は、本発明の実施例の構成を示す。図において、１は画像入力部、２は文字認識処理部、３は認識結果を文字種毎に測定する文字種測定部、４は入力文書画像とカテゴリインデクス内の代表画像との類似度を基にカテゴリを決定する文書画像カテゴリ決定部、５はカテゴリを代表する文書画像の文字種別頻度データを格納したカテゴリインデクス、６はカテゴリインデクス作成部、７は全ての原稿の文字種別頻度データを格納した頻度データ格納部、８は画像カテゴリ格納部、９、１０、１１は各カテゴリに分類された画像データである。

【００１４】図２は、本発明の処理フローチャートである。スキャナなどの画像入力部１から、文書などのイメージデータを読み込み（ステップ１０１）、文字認識処理部２では、読み込まれたデータについて文字認識処理を行い（ステップ１０２）、その認識結果を文字種測定部３に入力する。

【００１５】文字種測定部３は、上記した文字認識結果を、英字／数字／記号などの文字種ごとに頻度を測定する（ステップ１０３）。ここで、測定対象文字としては、認識結果の全ての文字を対象としてもよいし、認識結果の内、信頼度の高い文字のみを対象としてもよい。また、文字種測定部３では総文字数も測定する。

【００１６】文書画像カテゴリ決定部４は、測定した頻度データを基に、カテゴリインデクス５内の既存カテゴリのどれに分類可能かを決定する。ここで、カテゴリインデクス５には、各カテゴリの特徴を表すのに最も適した文書画像（代表画像）を１つ選択し、その代表画像の文字種別頻度データが格納されている。図３は、カテゴリインデクスの一例を示す。このカテゴリインデクスは、カテゴリインデクス作成部６によって、予め既定の画像が用意されているとき、もしくはカテゴリを再設定した際に自動的に作成される。この代表画像を求める方法としては種々の方法があるが、例えば、文書画像をある特徴空間にマッピングした際にグループ（カテゴリ）の中心に位置するものを代表画像として用いる。

【００１７】文書画像カテゴリ決定部４は、入力文書画像の文字頻度データと、カテゴリインデクス５に格納されている文字頻度データを対象として、文書画像の類似度を求める（ステップ１０４）。類似度の測定方法としては公知の手法を用いればよいが、主成分分析もしくは数量化理論第ＩＶ類などを用いるのが望ましい。最終的に、文書画像を座標上の空間にマッピングして距離の近いものどうしを同一カテゴリと定める。図４は、画像分類のマッピング例を示し、画像番号（０３２）がカテゴリＡに分類され、画像番号（０２３）がカテゴリＢに分類されている。これによって入力画像は、最も距離の近い文書画像と同一のカテゴリに分類される（ステップ１０５、１０６）。

【００１８】また、何れのカテゴリとも距離が遠い場合（距離が所定の閾値Ｔｈ以上）もある（ステップ１０５でＮｏ）。その場合は、文書画像カテゴリ決定部４はカテゴリインデクス作成部６に対してカテゴリの再設定を指示する。カテゴリインデクス作成部６は、頻度データ格納部７に格納されている全原稿の頻度データを対象として、文書画像間の類似度を求め、文書画像を再度分類する。さらに、各カテゴリの代表画像を再設定し、カテゴリインデクス５を再作成する（ステップ１０７）。図４の例では、例えばカテゴリＡがカテゴリＡ１とカテゴリＡ２に再設定される。

【００１９】ステップ１０７からステップ１０４に進み、文書画像カテゴリ決定部４は、再設定されたカテゴリインデクス５を参照して前述したと同様に類似度を求め（ステップ１０４）、入力文書画像のカテゴリを決定する。

【００２０】上記した文字種として、例えば数字または英字の頻度を用いると、文書画像群などの分類により効果的である。すなわち、例えば、数字の比率が高くかつ文字の量が多い文書画像を電話帳と分類し、数字の比率が高くかつ文字の量が少ない文書画像を帳票と分類し、さらに、英字の比率が高くかつ文字の量が多い文書画像を英文書と分類する。

【００２１】なお、本発明は上記したものに限定されず、ソフトウェアによっても実現することができる。本発明をソフトウェアによって実現する場合には、図５に示すように、ＣＰＵ、ＲＯＭ、ＲＡＭ、表示装置、ハードディスク、キーボード、ＣＤ－ＲＯＭドライブなどからなる汎用の処理装置を用意し、ＣＤ－ＲＯＭなどのコンピュータ記憶媒体には、本発明の文書画像分類機能を実現するプログラムが記録されている。

【００２２】

【発明の効果】以上、説明したように、請求項１記載の発明によれば、文書画像を分類する場合に文字種情報を用いているので、単語辞書を使用するキーワード検索などに比べて比較的簡単に該情報を得ることができ、高速に文書画像を分類することができる。また、文字認識処理は認識結果が必ずしも正確であるとは言えないが、本発明では文字自体の頻度ではなく文字種を測定しているので、多少の誤りがあっても精度に及ぼす影響が少なく、精度を落すことなく、文書画像を自動的に分類することができる。

【００２３】請求項２記載の発明によれば、文字種の特徴が類似している文書をカテゴリとすることで、文字種の特徴をもつ文書画像群を、高速に分類することができる。

【００２４】請求項３記載の発明によれば、カテゴリの代表画像を選択しているので、すべての文書について類

似度を測定する必要がなくなり、より高速に文書画像を分類することができる。

【００２５】請求項４記載の発明によれば、代表画像と適合しなかった場合にのみ、カテゴリを再設定しているので、より高速に分類することができるとともに、代表画像のみによる分類精度の低下も抑えることができる。

【００２６】請求項５記載の発明によれば、文字種の特徴として、数字または英字の頻度を用いているので、文書画像群を効率的に分類することができる。

【図面の簡単な説明】

【図１】本発明の実施例の構成を示す。

【図２】本発明の実施例の処理フローチャートを示す。

【図３】カテゴリインデクスの一例を示す。

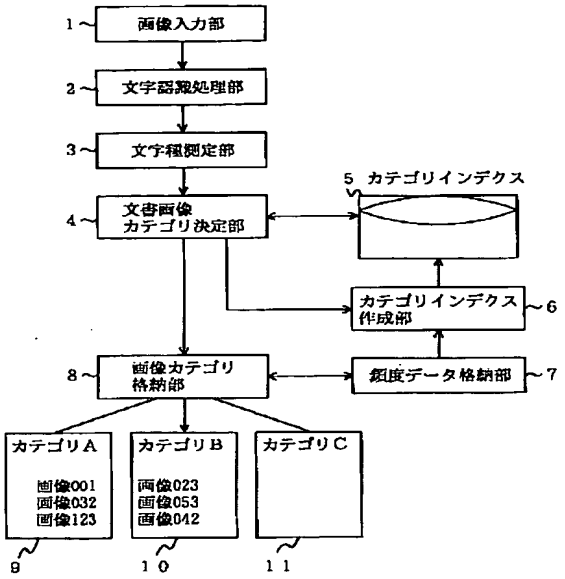
【図４】画像分類のマッピング例を示す。

【図５】本発明をソフトウェアによって実現する場合の構成例を示す。

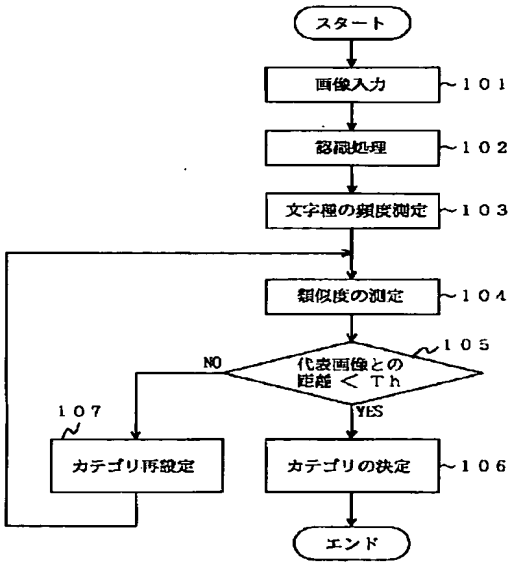
【符号の説明】

1　画像入力部
2　文字認識処理部
3　文字種測定部
4　文書画像カテゴリ決定部
5　カテゴリインデクス
6　カテゴリインデクス作成部
7　頻度データ格納部
8　画像カテゴリ格納部
9、１０、１１　各カテゴリに分類された画像データ

【図１】



【図２】



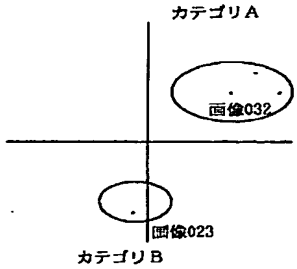
【図３】

カテゴリインデクスの例

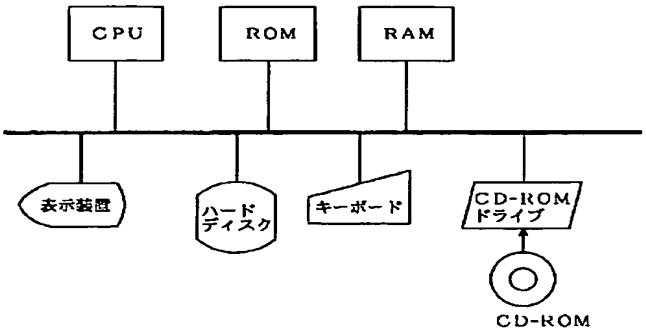
	数字	英字	記号	カナ	かな	漢字	総頻度	代表画像 No
カテゴリA	200	13	45	15	4	10	287	画像 032
カテゴリB	20	10	2	50	20	300	402	画像 023

【図４】

画像分類のマッピング例



【図５】



フロントページの続き

(72)発明者　幸地　司
東京都大田区中馬込１丁目３番６号　株式　20
会社リコー内